BMC Bioinformatics

BioMed Central

Software

GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction

Yao Yu†[1,2], Kang Tu†[1], Siyuan Zheng[1], Yun Li[1], Guohui Ding[1], Jie Ping[4], Pei Hao*[1,3] and Yixue Li*[1,2,3,4,5]

Address: [1]Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, PR China, [2]Graduate School of the Chinese Academy of Sciences, Shanghai 200031, PR China, [3]Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, PR China, [4]College of life science and biotechnology, Shanghai Jiaotong University, Shanghai 200240, PR China and [5]College of life science and biotechnology, Shanghai Tongji University, Shanghai 200331, PR China

Email: Yao Yu - yyu01@sibs.ac.cn; Kang Tu - ktu@sibs.ac.cn; Siyuan Zheng - syzhenger@gmail.com; Yun Li - yli01@sibs.ac.cn; Guohui Ding - gwding@sibs.ac.cn; Jie Ping - pjtalent@sjtu.edu.cn; Pei Hao* - phao@sibs.ac.cn; Yixue Li* - yxli@sibs.ac.cn

* Corresponding authors †Equal contributors

Published: 25 August 2009

BMC Bioinformatics 2009, **10**:264 doi:10.1186/1471-2105-10-264

Received: 13 May 2009
Accepted: 25 August 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/264

Abstract

Background: In the post-genomic era, the development of high-throughput gene expression detection technology provides huge amounts of experimental data, which challenges the traditional pipelines for data processing and analyzing in scientific researches.

Results: In our work, we integrated gene expression information from Gene Expression Omnibus (GEO), biomedical ontology from Medical Subject Headings (MeSH) and signaling pathway knowledge from sigPathway entries to develop a context mining tool for gene expression analysis – GEOGLE. GEOGLE offers a rapid and convenient way for searching relevant experimental datasets, pathways and biological terms according to multiple types of queries: including biomedical vocabularies, GDS IDs, gene IDs, pathway names and signature list. Moreover, GEOGLE summarizes the signature genes from a subset of GDSes and estimates the correlation between gene expression and the phenotypic distinction with an integrated p value.

Conclusion: This approach performing global searching of expression data may expand the traditional way of collecting heterogeneous gene expression experiment data. GEOGLE is a novel tool that provides researchers a quantitative way to understand the correlation between gene expression and phenotypic distinction through meta-analysis of gene expression datasets from different experiments, as well as the biological meaning behind. The web site and user guide of GEOGLE are available at: http://omics.biosino.org:14000/kweb/workflow.jsp?id=00020

Background

The rapid development of high-throughput gene expression detection technology provides a huge amount of experimental data for advanced researches on associating gene expression signatures with biological phenotypes. The application of microarrays to identify gene expression signatures of human diseases has been widely accepted [1,2]. Accordingly, a vast number of tools for microarray analysis are available, including ArrayPipe [3], GEPAS [4], GeneTrailExpress [5], and currently reported Perl modules for microarray analysis [6], etc. Besides, Gene Set Analysis is highlighted in microarray analysis. Gene sets are usually defined as set of genes which function in cohort, detailed analysis on which can lead to a functional level map of the transcriptome data. Some popular gene set analysis tools include Babelomics [7], WebGestalt [8], etc. Furthermore, to address the problems of limited samples in single biological experiment and heterogeneity of gene expression datasets from different sources, methods for large-scale meta-analysis of microarray data have been developed [9-11]. Those tools for meta-analysis like studies such as Connectivity Map [12] requires a huge amount of supporting data resources, and associated information from existing biological databases. There is a clear requirement for efficiently retrieving associative datasets for meta-analysis to avoid manual work in mining from a large number of references.

The Gene Expression Omnibus (GEO) [13], curated by the National Center for Biotechnology Information (NCBI), is designed in response to this demand as a public warehouse for the submission, storage and retrieval of the high-throughput gene expression and genomic hybridization experiments. Several tools and strategies for operating the GEO database have been developed to enable comparisons of microarray data across experimental platforms, different laboratories and multiple species [14-18]. However most of these tools for retrieving data from the GEO repository paid little attention to mining further information about the gene expression signatures, such as linking to the biological functions of genes, or integrating the related pathway information in the biological processes. The National Library of Medicine's controlled vocabulary thesaurus (MeSH) [19] is one of the best resources for biomedical vocabularies. MeSH is helpful to be used as an index to link experimental conditions and biological concepts together, including disease phenotypes.

Focusing on this issue, we developed a state-of-the-art online bioinformatics tool, named GEOGLE, for mining the experimental data from GEO database and constructing relationships among the datasets, genes, pathways and the genes' biological significance. Our system integrates information from multiple sources, such as sigPathway [20] for pathway information and MeSH for biomedical vocabularies. Investigators are able to use multiple types of data for querying – including disease information, gene symbols, pathway names, expression datasets (GDS IDs), and signature lists – to search a large collection of related microarray information. An integrated p value is introduced by GEOGLE, which could be considered as an estimate for the correlation between gene expression and the phenotypic distinction. This mining technology may have great value in discovering the linkages between known phenotypes and experiment data, as well as retrieving suitable datasets for further research work.

Implementation

GEOGLE description and results

The analysis in GEOGLE consists of two major parts: meta-analysis which integrates literature information and similarity search for signatures and datasets. The gene expression data and the basic signature for each dataset are derived from public expression data warehouses, such as GEO. MeSH terms have been used as important vocabulary dictionary for associating gene expression data with other biological terms, such as pathways and diseases. Dataset searching is mainly based on checking synonyms from pathways or diseases' description of their relevant MeSH terms in dataset annotation. Through literature searching and dataset filtering, summarized signatures are available from the integration of GEO, MeSH and sigPathway. For the second part of signature similarity search, a similar method has been used in Connectivity Map [12]. GEOGLE will search similar datasets sharing the same signatures from the databases. By associating the attributes of these datasets, GEOGLE will summarize the common features and suggest the potential relationships between genes and diseases.

Datasets collection and signature extraction

An architectural pattern was used in the design of GEOGLE that isolated data processing logic from user interface considerations. Each part of this pattern is independent for more convenient maintenance (Fig. 1). The first module is a data collection and signature extraction engine. Currently, gene expression datasets collected in GEOGLE mainly consist of data in three species from GEO: Human, Mouse and Rat. In total 1005 GEO datasets (GDSes) from

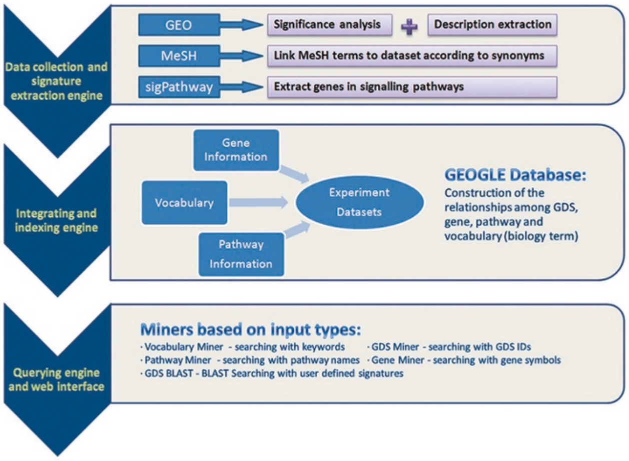
Data collection and signature extraction engine
GEO
Significance analysis
Description extraction
MeSH
Link MeSH terms to dataset according to synonyms
sigPathway
Extract genes in signalling pathways
Integrating and indexing engine
Gene Information
Vocabulary
Pathway Information
Experiment Datasets
GEOGLE Database:
Construction of the relationships among GDS, gene, pathway and vocabulary (biology term)
Querying engine and web interface
Miners based on input types:
· Vocabulary Miner - searching with keywords
· GDS Miner - searching with GDS IDs
· Pathway Miner - searching with pathway names
· Gene Miner - searching with gene symbols
· GDS BLAST - BLAST Searching with user defined signatures

Figure 1

The components of GEOGLE. GEOGLE contains three major engines: the data collection and signature extraction engine, the integrating and indexing engine, the querying engine and web interface.

21 platforms (GPLs) have been collected in GEOGLE. 6 GPLs in Human contain 351 GDSes, while 9 GPLs in Mouse contain 500 GDSes and 6 GPLs in Rat contain 154 GDSes. The details of these datasets are available from our web site. Some GEO datasets will be filtered out because of the limited chip numbers (i.e. less than 3 samples in each subset), which might be considered not suitable for meta-analysis. Each dataset derived was classified into different groups according to corresponding experimental factors (e.g. tissue, strain, time, dose, etc). For example, GDS1436 (Cigarette smoking effect on alveolar macrophages) contains 10 samples. According to GSMs' labels these samples can be classified into two groups: non-smoker and smoker group. After classification, differential expression of each gene in each data set is tested using Significance Analysis of Microarray (SAM). SAM is a statistical technique for finding differentially expressed genes in microarray experiments http://www-stat.stanford.edu/~tibs/SAM/. The p values from SAM reflect the correlation between gene expression and the phenotypic distinction. (A detail description of SAM should be found in the supplementary file) Biomedical vocabulary information was collected from MeSH. Considering the hierarchical structure of MeSH, we automatically associated the corresponding MeSH terms to all its stored synonyms. This engine can be also used to update GEOGLE by collecting data from web sources.

Data integration and indexing

These data are processed by a second integrating and indexing engine. Three kinds of relationships were constructed through this engine and stored in the GEOGLE database, such as the linkage between gene and experiment dataset, between pathway and dataset, and between vocabulary and dataset. The linkage between gene and experiment dataset was represented in two aspects: the individual p value for estimating the significance of differential gene expression in one dataset, and the integrated p value for estimating the correlation between gene expression and the phenotypic distinction which might contain several datasets with similar phenotypes. The algorithm for calculating the integrated p value is presented in the next paragraph. To construct the linkage between pathway and dataset, signatures from datasets were mapped into pathways based on information from sigPathway dataset. MeSH terms were mapped into experiments' annotation according to the context mining of the recorded synonyms from the description of GDSes.

The procedure for calculating the integrated p value consists of two major parts. Firstly, the p value of each gene in individual dataset was calculated with SAM method, as mentioned before. Secondly, a novel procedure was developed to calculate an integrated p value for evaluating the relationship between signature and a group of referred datasets (reflecting a phenotype). Steps from (1) to (5) were performed:

(1) The p values of different genes in each dataset were organized into a vector. Then gene – GDS matrices (named P_{gc}, *gc* for gene – condition) were generated from a set of p value vectors calculated independently from different GDSes. Each element in P_{gc} represents a p value which had been prepared before using SAM.

(2) What we want to know is if these genes in P_{gc} are perturbed under a group of GDSes, which equal to test if the sub matrix of P_{gc} of those genes and those GDSes (named P_{gc_sub}) follow uniform distribution. To perform such a test, P_{gc_sub} was transformed to Z_{gc_sub} with quantile function of normal distribution. A quantile function of a probability distribution is the inverse F^{-1} of its cumulative distribution function (cdf) F. Assuming a continuous and strictly monotonic distribution function, $F:R$ -> (0,1), the quantile function returns the value below which random draws from the given distribution would fall, $p \times 100$ percent of the time. That is, it returns the value of x such that

$$\Pr(X \le x) = p$$

The cumulative distribution function (cdf) of the normal distribution is expressed in terms of the density function as follows:

$$\varphi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \phi_{\mu,\sigma^2}(u)du = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)du$$

$$x \in \Re$$

(3) Then Z score was summarized from Z_{gc_sub} with the function:

$$Z = \frac{sum(Z_{gc\_sub})}{\sqrt{n}}, \; n \text{ for the number of elements in } Z_{gc\_sub}$$

(4) A new p value was calculated to represent the significance of Z score using the cumulative distribution function of normal distribution (as mentioned in (2)). Let a parameter ('alpha') be the threshold of the p value from this test.

(5) If these genes were not signatures of a group of GDSes, the P_{gc_sub} would follow uniform distribution. If P_{gc_sub} followed uniform distribution, Z_{gc_sub} would follow norm distribution. As a result, Z score would also follow norm distribution. A significant small value of Z comparing to normal distribution corresponded to the significantly being perturbed of these genes under these conditions. The p value from this test is considered as the integrated p value from the whole searching task. We could judge

whether certain gene should be considered as signature in the group of GDSes by the integrated p value.

(6) To judging the relationship between candidate signatures and vocabularies is very similar with the procedure from (1) to (5) mentioned before. Each vocabulary (MeSH term) contains a groups of expression datasets (GDSes). The integrated p value for the significance of the correlation between signature and certain vocabulary is equal to that for the correlation between signature and the GDSes in the vocabulary.

(7) The next step is to evaluate the relationship between pathway and a group of genes (signatures). We used a very similar procedure with some modification. We constructed a pathway – gene matrices (named P_{pg}, *pg* for pathway – gene). The relationship of pathways and genes were derived from sigPathway. Each element in this matrix is the integrated p value of gene in a group of GDSes (this group is determined according pervious GDS searching). Then the procedure from (1) to (5) was repeated, using P_{pg} taking the place of P_{gc} as initial input. The new integrated p values calculated were considered to be the estimate of the significance of the pathways in the searching task.

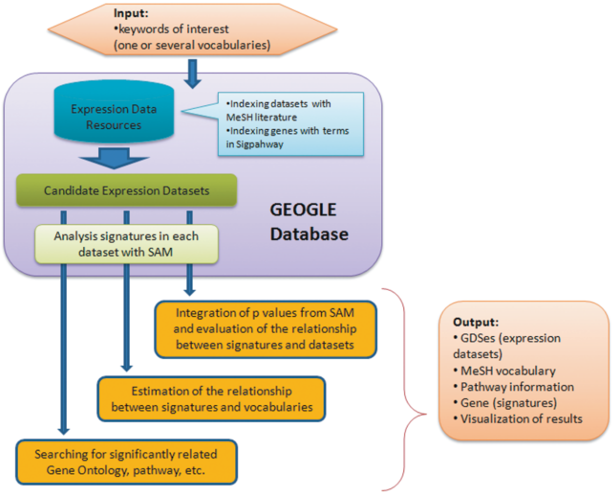
Input:
•keywords of interest
(one or several vocabularies)
Expression Data Resources
•Indexing datasets with MeSH literature
•Indexing genes with terms in Sigpahway
Candidate Expression Datasets
GEOGLE Database
Analysis signatures in each dataset with SAM
Integration of p values from SAM and evaluation of the relationship between signatures and datasets
Estimation of the relationship between signatures and vocabularies
Searching for significantly related Gene Ontology, pathway, etc.
Output:
• GDSes (expression datasets)
• MeSH vocabulary
• Pathway information
• Gene (signatures)
• Visualization of results

Figure 2

The procedure for Vocabulary Miner. This figure illustrates the main process in performing Vocabulary Miner. User can input certain keyword(s) for searching. GOEGLE will give the results contain: GDSes (expression datasets), significantly related MeSH vocabulary, involved pathway information, genes (signatures) and some visualization of results.

Results

The third part of the pattern is a querying engine where users can perform searching tasks with a friendly web interface. The implementation of this engine is based on workflow technique using R and Java for script language and Omics Explorer (manuscript under review) as a web container. There are five miners provided by GEOGLE. The main difference of these miners includes various input types for querying including biomedical vocabularies, GDS identifies, pathway names, gene identifies or a set of signatures, and slightly different searching strategies. **Vocabulary Miner** (Fig. 2): Mining out relevant GDSes, common signature genes and pathways based on biomedical vocabularies. The submitted keywords (one or several vocabularies) of users' interest will be mapped to a series of related GDS according to MeSH dictionary and the description of GDS. The integrated P value which describes the relationship between each gene and the group of returned GDSes (as well as MeSH terms) is calculated. Meanwhile, the P value which describes the relationship between each pathway and the group of GDSes is calculated. Then significantly associated GDSes, genes and pathways will be individually sorted based on the p values and be given out with necessary annotation. **GDS Miner**: This miner is similar with the first one. However the datasets (usually GDS IDs) should be directly given by the users instead of being found by searching keywords. The following process of retrieving signature genes and extracting related pathway information is the same as Vocabulary Miner. Also this miner can be helpful in searching for the annotation information of datasets with GDS IDs. **Pathway Miner**: Users may submit certain pathway name or partial name. This miner will help to obtain those genes within the pathway. Then the GDSes in which these genes are considered as signatures will be returned as well as their annotation. Similarly, GEOGLE will summarize the correlation between pathways (or genes in the pathways) and different datasets. **Gene Miner**: Users can provide several genes of their interest directly as input. Then the miner will search for the associative GDSes in which these genes are considered to be signatures. The remaining process is quite similar with Pathway Miner. **GDS Blast Miner**: Users may submit a list of signatures according to their own definition. What this miner will perform is to search for similar datasets containing the same (or similar) group of signatures. Then this miner will summarize associated biomedical vocabularies to these GDSes. The format of submitted datasets is very flexible, examples of which could be a gene list with expression values or probe set IDs with FDR values (see Additional file 1).

Some optional parameters can be set by users. For instance, in Vocabulary Miner setting 'F' (false) for 'list_Mesh_GDS.only' makes the miner search additional information about genes and pathways according to the query. A more efficient searching, by setting 'T' for this parameter, comes at the cost of performing no pathway and gene information searching. Another common parameter is 'alpha'. This parameter sets the threshold for the integrated p value to estimate the correlation between gene expression and experiment datasets. GEOGLE provides a task management system for users to review the states of their previously conducting tasks and to retrieve the results later, which will be saved temporally on the server. The detail processing pipelines of these miners and a step-by-step tutorial of using GEOGLE and the explanation of the input and results could be found in the Supplementary.

By using 'Vocabulary Miner', we search for 'smoking' related gene expression gene data in human, then got a result with 4 GDSes considered to be candidate datasets (GDS1304, GDS1436, GDS1673 and GDS534). According to their annotation, these GDSes which are all related to cigarette smoking effect are suitable for further meta-analysis. Terms like 'Breast Cancer/Estrogen Receptor Signaling' and 'Stress Response to Cellular Damage' are returned with significant p values in pathway section of the results, which suggests that these pathways are closely related to the 'smoking' phenotype. Such genes like GALNT1 are identified as signature genes trough all these dataset. According to pervious report, GALNT1 is strongly associated with the using of tobacco and the risk of lung cancer [21,22]. In our previous work, GEOGLE severed as main tool for expression data analysis associated with metabolomics data, which reveals distinct variations related to nicotine consumption in human [23]. The combination of several miners provided not only suitable expression datasets but also candidate genes which might be related to the influence of smoking. The gene for alkylglycerone phosphate synthase (*alkyl-DHAP*, or *AGPS*) has been found strong down-regulated in smokers in human lung tissues. This is consistent with metabolic profiling. The down-regulation of this gene was found to influence both ether lipid and glycerophospholipid pathways, and shift the ratios of plasmalogens to diacyl-phosphatidylcolines.

Conclusion

In this report we introduce GEOGLE, an online web service for GEO dataset mining and biomedical information integration. GEOGLE provides an efficient way for users to search for related experiment datasets according to their own research interest with various types of input. Another significant feature of GEOGLE is the novel concept of an integrating system for signatures, pathways, biological terms and disease information. Public data warehouses such as GEO are high-quality resources for an automatic mining and integration system of gene expression datasets and reference literatures from GEOGLE, which will be a revolution compared to manually collecting experimental data for biological research. Currently there exist a few tools for operating the GEO database. For example, Oncomine [24,25] is a previously published cancer gene expression analysis platform. CleanEx [26] also contains re-annotate experiment datasets with the MeSH terms and some on-line analysis tools for gene expression data. Compared with these tools, GEOGLE has some outstanding features and additional values for this kind of study. Firstly, the main object of GEOGLE is to search for candidate datasets from different experiments for further meta-analysis, according to certain biological vocabularies and/or genes of interest. Secondly, GEOGLE provides a quantitative method to evaluate the correlation between each gene and a series of gene expression datasets which might represents certain phenotypic distinction. Thirdly, GEOGLE collected a wide range of information about different kinds of diseases including cancer (over 60,000 MeSH terms have been involved). Fourthly, GEOGLE performed further mining for related gene function information, pathway annotation and reference knowledge and introduced an integrated p value for estimating the correlation between gene expression and the phenotypic distinction. Fifthly, GEOGLE allows multiple types of inputs such as keywords, datasets, pathways, genes and user defined signatures. Technically, a modular design concept allows each part of GEOGLE to be replaced by a more advanced one, for instance another BLAST engine with more accuracy could be used for the similarity search. The container of GEOGLE (Omics Explorer) is hosted via a standard online service platform supported by InforSense Ltd. Thus no individual GUI will be need for GEOGLE's online user interface. In addition, the GEOGLE database can be easily updated to keep it synchronized with public gene expression databases.

Further steps in the development of GEOGLE should focus on the integration of high-throughput gene expression databases other than GEO, such as the ArrayExpress [27] and the Stanford Microarray Database (SMD) [28]. One of the improvements of GEOGLE in-progress is large scale gene and disease information mining effort from reference databases [29] and integrating this information with existing signature data. The reference mining results are believed to be able to prove the reliability of the relationships between signatures and diseases discovered by GEOGLE. Moreover, since GEOGLE provides a potential network of diseases, genes and pathways, more analysis work focusing on this will be considered in future.

Availability and requirements

- **Project name**: GEOGLE

- **Project home page**: http:omics.biosino.org:14000/kweb/work flow.jsp?id=00020

- **Operating system(s)**: Developed in Linux and platform independent for accessing

- **Programming language**: Java 1.5 and R 2.5.1

- **Other requirements**: Internet Explorer, Firefox or Safari is required to access the website.

Abbreviations

GEO: Gene Expression Omnibus; MeSH: Medical Subject Headings; SAM: Significance Analysis of Microarray; NCBI: National Center for Biotechnology Information.

Authors' contributions

YY and KT made substantial contributions to conception and design of GEOGLE as well as in preparing the manuscript. SZ and GD also participated in the construction of GEOGLE and gave important suggestion to the manuscript. JP was involved in the maintaining of the web server. PH and YL participated in the design of the study and preparing of the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

*GEOGLE – **Supplementary Material**. The supplementary materials of GEOGLE for user manual and the description of methods.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-264-S1.doc]

Acknowledgements

Funding: The 863 Hi-Tech Program of China (863) (grant 2007AA02Z304, 2006AA020406), the Shanghai Committee of Science and Technology (Grant 07dz22004, 08JC1416600) and Research Program of CAS (grant KSCX2-YW-R-112).

References

1. Glinsky GV, Berezovska O, Glinskii AB: **Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.** *J Clin Invest* 2005, **115(6):**1503-1521.
2. Vijver MJ van de, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, *et al.*: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347(25):**1999-2009.
3. Hokamp K, Roche FM, Acab M, Rousseau ME, Kuo B, Goode D, Aeschliman D, Bryan J, Babiuk LA, Hancock RE, *et al.*: **ArrayPipe: a flexible processing pipeline for microarray data.** *Nucleic Acids Res* 2004:W457-459.
4. Tarraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcarate S, Goetz S, Escobar P, *et al.*: **GEPAS, a web-based tool for microarray data analysis and interpretation.** *Nucleic Acids Res* 2008:W308-314.
5. Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O, Kaufmann M, Lenhof HP: **GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments.** *BMC Bioinformatics* 2008, **9:**552.
6. Morris JA, Gayther SA, Jacobs IJ, Jones C: **A suite of Perl modules for handling microarray data.** *Bioinformatics* 2008, **24(8):**1102-1103.
7. Al-Shahrour F, Carbonell J, Minguez P, Goetz S, Conesa A, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J: **Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments.** *Nucleic Acids Res* 2008:W341-346.
8. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005:W741-748.
9. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101(25):**9309-9314.
10. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res* 2002, **62(15):**4427-4433.
11. Hao P, Zheng S, Ping J, Tu K, Gieger C, Wang-Sattler R, Zhong Y, Li Y: **Human gene expression sensitivity according to large scale meta-analysis.** *BMC Bioinformatics* 2009, **10(Suppl 1):**S56.
12. Lamb J: **The Connectivity Map: a new tool for biomedical research.** *Nat Rev Cancer* 2007, **7(1):**54-60.
13. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30(1):**207-210.
14. Barrett T, Edgar R: **Gene expression omnibus: microarray data storage, submission, retrieval, and analysis.** *Methods Enzymol* 2006, **411:**352-369.
15. Barrett T, Edgar R: **Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*.** *Methods Mol Biol* 2006, **338:**175-190.
16. Boyle J: **Gene-Expression Omnibus integration and clustering tools in SeqExpress.** *Bioinformatics* 2005, **21(10):**2550-2551.
17. Sean D, Meltzer PS: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.** *Bioinformatics* 2007, **23(14):**1846-1847.
18. Yi Y, Li C, Miller C, George AL Jr: **Strategy for encoding and comparison of gene expression signatures.** *Genome Biol* 2007, **8(7):**R133.
19. Lowe HJ, Barnett GO: **Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches.** *JAMA* 1994, **271(14):**1103-1108.
20. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102(38):**13544-13549.
21. Sellers TA, Huang Y, Cunningham J, Goode EL, Sutphen R, Vierkant RA, Kelemen LE, Fredericksen ZS, Liebow M, Pankratz VS, *et al.*: **Association of single nucleotide polymorphisms in glycosylation genes with risk of epithelial ovarian cancer.** *Cancer Epidemiol Biomarkers Prev* 2008, **17(2):**397-404.
22. Wright JM, Merlo CA, Reynolds JB, Zeitlin PL, Garcia JG, Guggino WB, Boyle MP: **Respiratory epithelial gene expression in patients with mild and severe cystic fibrosis lung disease.** *Am J Respir Cell Mol Biol* 2006, **35(3):**327-336.
23. Wang-Sattler R, Yu Y, Mittelstrass K, Lattka E, Altmaier E, Gieger C, Ladwig KH, Dahmen N, Weinberger KM, Hao P, *et al.*: **Metabolic profiling reveals distinct variations linked to nicotine consumption in humans–first results from the KORA study.** *PLoS One* 2008, **3(12):**e3863.
24. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform.** *Neoplasia* 2004, **6(1):**1-6.
25. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, *et al.*: **Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.** *Neoplasia* 2007, **9(2):**166-180.
26. Praz V, Bucher P: **CleanEx: new data extraction and merging tools based on MeSH term annotation.** *Nucleic Acids Res* 2009:D880-884.
27. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, *et al.*:

ArrayExpress: a public database of gene expression data at EBI. *C R Biol* 2003, **326(10–11):**1075-1078.

28. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, *et al.*: **The Stanford Microarray Database: implementation of new analysis tools and open source release of software.** *Nucleic Acids Res* 2007:D766-770.
29. Zheng S, Sheng J, Wang C, Wang X, Yu Y, Li Y, Michie A, Dai J, Zhong Y, Hao P, *et al.*: **MPSQ: a web tool for protein-state searching.** *Bioinformatics* 2008, **24(20):**2412-2413.